Final Project

Vicki Hertzberg

January 21, 2017

Final Project

Over the course of this semester you will work on a data science project. The goal of this assignment is to have you proceed throught the steps of a data science project as we outlined on the first day of class, from creating a question and acquiring data, through data wrangling, statistical analysis, and visualization, to production of a publication. In this case you will be identifying a data set of interest to you that you can use to address a question of interest to you. You will obtain the dataset, reshape it so that it fits your analytic plan, clean it as necessary, explore it, analyze it, create visualizations of your results, and write a manuscript describing what you did and what you found. You will also create a public website for your project.

Milestones

There are two intermediate milestones and three final milestones for your project. Extensions of the due date will not be given. If you anticipate an issues (e.g., business travel), you must send Vicki and Melinda an email at least one week in advance.

Date	**Description**
Wednesday February 1 by **5:00 pm EST**	Submit a project proposal (Milestone 1)
Wednesday 15 March by **5:00 pm EDT**	Submit a working prototype (Milestone 2)
Wednesday 19 April by **5:00 pm EDT**	Submit the final PDF of your manuscript (Milestone 3)
Wednesday 26 April by **5:00 pm EDT**	Submit the file contining your presentation material

Deliverables

You will also create a website for your project using GitHub. This website should display a summary of the main results of your project, telling its story. Embed your main visualizations on this site. Please write this so that your grandmother could understand it. Your Rmd file, HTML file, and dataset should be linked from your GitHub repository.

Project Milestone 1 - Proposal

Your project proposal, Milestone 1, is due on Wednesday, 1 Februray 2017, by **5:00 pm EST**.You will submit a PDF file. This document will contain two links; 1. a link to the Rmd file that created it in your GitHub repo, and 2. a link to the source of your original data. Please put your name on the first page of the document. The document will include the following information:

- Basic information: Project title, your name, email address
- Overview and Motivation: Why did you undertake this particular project? What inspired you, what are your background and research interests that may have influenced your decision?
- Project Objectives: What is the primary focal question that you are trying to answer? What would you like to learn and accomplish?
- Data: from where and how are you acquiring your data? **Provide a link to your data source.**
- Data Wrangling: Do you anticipate that there will be extensive data cleaning / reshaping / extraction? Are there questions you will need to calculate in your data (e.g., perhaps you have height and weight, but not BMI)? How will you implement this particular data wrangling step?
- Exploratory Analysis: Which methods / visualizations are you planning to use to explore your tidy dataset?
- Analysis: How are you planning to analyze your data? _ Schedule, keeping in mind the due dates listed above for the intermediate and final milestones, make a plan to meet these deadlines. Write these in terms of weekly tasks / goals.

As a ballpark, your proposal should be about 2-3 pages of text, tables, and figures. You could even include some preliminary data acquisition / analysis steps.

After we receive your proposals we will find a time to meet with you to discuss your proposal and also to help guide you through the rest of the analysis.

Project Milestone 2 - A Working Prototype

A working prototype of your project is due on Wednesday 15 March by **5:00 pm EDT**. You will submit a PDF file. This document will contain two links; 1. a link to the Rmd file that created it in your GitHub repo, and 2. a link to the source of your original data. Please put your name on the first page of the document.

For this milestone we expect that you will have acquired, cleaned, and explored your dataset. You will document these activities in your files. You will also explain in detail the components of your final analysis. If you have deviated from your original plans, please describe what is different and the reasons that have led to this. You will explain your workflow, that is, how you went from acquiring your data to getting (close to) an answer to your question. Are there ancillary questions that have arisen as you have gone through this process? What are they?

As a ballpark number, this document should be about 10 pages of text, tables, and figures.

Project Milestone 3 - Your Final Document

You will complete the analyses you described in your prototype, and write a manuscript. You will submit a PDF file. This document will contain two links; 1. a link to the Rmd file that created it in your GitHub repo, and 2. a link to the source of your original data. Please put your name on the first page of the document.

You will formulate this document as a manuscript, with abstract, main body, figures and tables.

Presentation

You will present your project in class on 26 April. You will each have 5 minutes to walk us through your project. We will strictly enforce the 5 minute time limit, so please rehearse thoroughly before hand. Use principles of good story telling and presentations to make your key points. Focus the presentation on your results and findings rather than technical details. What is the single most important thing you want your audience to remember? Make this point front and center. What insights did you gain by doing this project? What was the best part of the project?

Rmd and HTML Files

The Rmd and html files are important aspects of your project. They will detail your steps in developing your final paper and presentations, from your initial question, through how you obtained your data, the statistical methods you used and alternatives you considered. These files should include many visualizations.

Code

We expect that you will write high quality, readable R code in your Rmd files. You should aim to follow the processes we describe early in the class, thinking about things such as reproducibility, data cleaning, etc. We also expect you to document your code.

Submission of Milestone Documents and Final Presentation

You will submit these to the links in the class Canvas site.

Grading

Each of the milestones will be graded independently. The points are listed in the syllabus.

Elements

- Project Scope: Did you choose the appropriate complexity and level of difficulty for your project?
- Process: Did you follow the data science process and is it well documented in your GitHub repo?
    - *Question*

- Do you specify the type of question (e.g., exploratory, association, causality)?
- Was the question specified prior to data acquisition?
- Is there a context for the question?
- How will you know when you have answered the question, by what metric?

- *Experiment*

- Are the experimental design details recorded?
- Could the experiment be addressed with existing data?

- *Data checking*

- Univariate / multivariate summaries used?
- Outliers evident?
- What is the code for a missing value?

- *Tidy data*

- Each variable is 1 column.
- Each observation is 1 row.
- Are there data of different types in the same dataset?
- Is there a record of how you went from raw to tidy data?
- Is there a code book or data dictionary?
- Are all parameters / functions / units that are used identified?

- *Exploring the data*

- How many missing values are present?
- Univariate plots of data?
- Correlations explored with scatterplots?
- Are all data points in the correct range?
- Is ther an attempt to identify errors or miscoding?
- Should plots made on a log scale?

- *Inference*

- Do you provide estimates of uncertainty along with your results?
- Do you specify the larger population to which you wish to generalize your results?
- Have you identified potential confounders or effect modifiers?
- Have you identified and modeled correlations of your measurements of interest with time and/or space?

- *Prediction*

- Did you first split your dataset into training and validation sets?
- Did you apply cross validation, resampling, bootstrapping first to the training set?
- Did you create features first with the training set?
- Did you estimate parameters first with the training set?
- Did you fix all features, parameters, and models before applying them to the validation set?
- Did you apply only one model to the valudation data and report and error rate?

- *Causality*

> - Was there a randomization mechanism to assign group membership?
> - Are there reasons why causality migh not be appropriate (e.g., confounding, effect modification)?
> - If so, is there avoidance of inappropriate use of causal language?

- Results: Are your results correct in how they address your questions?
- Implementation: Is your code of good quality? Is it well documented? Is it appropriately polished, robust, reliable?
    - Did you avoid manual calculations?
    - Are there scripts that reproduce all results?
    - Has somebody else been able to independently run the scripts and verifie all results?
    - Have you saved the raw and tidy versions of the data?
    - Have you recorded all hardward and software versions used?
- Publication and Presentation: Are these clear, engaging, and effective? Do you tell the story well in each of these? Do you use appropriate visualizations to communicate your data?
    - Did you lead your presentations (written and oral) with a brief, clear statement of your question?
    - Did you explain the data source, measurement technology, and experimental design before you explain the model?
    - Did you specify the data analytic question?
    - Did you specify the exact model?
    - Did you explain the features used before the model?
    - Does each figure communicate information effectively and address the question of interest?
    - Do your figures have plain language axis labels and titles?
    - Is the font size for your figures sufficiently large?
    - Are there detailed captions for your figures such that taken together the figures (and tables) tell the story?

Rubric

Exemplary: The student provides details about background and explanation, including appropriate tables and figures, demonstrating knowledge, making appropriate conclustions, identifying gaps, while avoiding meaningless details.

Accomplished: The student gives some background and explanation with tables and figures.

Developing: The student provides descriptive information.

Beginning: The student provides the barest details.

Fail: The student provides no information.